mathematics

MDPI

Article

Formal Group Fairness and Accuracy in Automated Decision Making

Anna Langenberg [1], Shih-Chi Ma [1,2], Tatiana Ermakova [3] and Benjamin Fabian [1,2,*]

[1] Information Systems, Humboldt-Universität zu Berlin, 10178 Berlin, Germany; shih-chi.ma@hu-berlin.de or shih-chi.ma@th-wildau.de (S.-C.M.)
[2] EDIH pro_digital, Technical University of Applied Sciences Wildau, 15745 Wildau, Germany
[3] School of Computing, Communication and Business, Hochschule für Technik und Wirtschaft, University of Applied Sciences for Engineering and Economics, 10318 Berlin, Germany; tatiana.ermakova@htw-berlin.de
* Correspondence: benjamin.fabian@th-wildau.de

Abstract: Most research on fairness in Machine Learning assumes the relationship between fairness and accuracy to be a trade-off, with an increase in fairness leading to an unavoidable loss of accuracy. In this study, several approaches for fair Machine Learning are studied to experimentally analyze the relationship between accuracy and group fairness. The results indicated that group fairness and accuracy may even benefit each other, which emphasizes the importance of selecting appropriate measures for performance evaluation. This work provides a foundation for further studies on the adequate objectives of Machine Learning in the context of fair automated decision making.

Keywords: AI; machine learning; automated decision making; algorithmic bias; metrics; group fairness

MSC: 68T01

1. Introduction

Automated decision making based on Machine Learning (ML) has become an integral part of human lives, applied in several areas like credit lending, hiring, risk assessment or health care [1]. ML algorithms are adopted to make decisions on behalf of humans, which rely on past data rather than human judgement. They rely on statistical concepts and are, therefore, assumed to achieve more accurate and fair decisions compared to human assessment. However, decisions made by machines are actually not free of discrimination [2]. Therefore, achieving fair outcomes in addition to correct predictions from the ML model is essential.

There is no consensus in research on the compatibility of these two terms. According to [3], the majority of research claims that an increase in fairness comes at an unavoidable loss of accuracy [4–8]. In addition, [9] further explored "impossibility theorems" of several fairness metrics and accuracy. In contrast, some researchers highlight that the assumed trade-off is not necessarily a given [8,10–15]. Based on the contradicting findings, this work is dedicated to the following research question: What is the relationship between accuracy and fairness in the context of automated decision making?

In our study, we operationalize the term accuracy by various established formal metrics, such as Accuracy (ACC), Balanced Accuracy (BACC) score and F1 score, to assess the correctness of outcomes [16]. Fairness in the context of ML has no universal formal definition [17]. A lot of the metrics proposed by research claim to make fairness accessible by measuring the difference in outcomes between individuals or groups with respect to their protected features, such as race or gender [18,19]. For fairness, this study focused on group fairness metrics, Statistical Parity Difference (SPD) and Average Odds Difference (AOD), to assess the disparities between privileged and unprivileged groups based on their protected features [20], such as race or gender [18,19].

To answer the research question, several approaches are examined: first of all, baseline classifiers are trained on the two predefined data sets, Correctional Offender Management

Citation: Langenberg, A.; Ma, S.-C.; Ermakova, T.; Fabian, B. Formal Group Fairness and Accuracy in Automated Decision Making. *Mathematics* **2023**, *11*, 1771. https://doi.org/10.3390/math11081771

Academic Editor: Alexander Ryzhov

Received: 10 March 2023
Revised: 31 March 2023
Accepted: 4 April 2023
Published: 7 April 2023

Profiling for Alternative Sanctions (COMPAS) and Law School Admissions Council's National Bar Passage Study (LAW), to ensure comparability among the analyzed approaches. A Logistic Regression Classifier (LRC) is optimized with respect to the two objectives. First, accuracy is maximized to analyze the effect on fairness. Ref. [12] highlights that optimizing a model with respect to accuracy decreases the numbers of errors made by the model and, hence, increases fairness in outcomes.

Second, by applying the popular pre-processing strategy Reweighing (RW), fairness is improved by mitigating biases [21].

In addition, general ML practices expected to increase accuracy are studied with respect to their effects on fairness in terms of the following: the size of the feature set and of the training set. Ref. [11] argued that the number of features used to make a decision has a positive impact on fairness, whereas the effect achieved by increasing the training set size is negative due to inherent data biases. This study verifies the effects of these practices on the studied datasets.

Finally, the data augmentation technique proposed by [10] is replicated to explore the accuracy–fairness relationship. Ref. [10] showed that augmenting synthetic data, representing an ideal world, to diminish the effect of the protected feature on the label, increases fairness and accuracy, simultaneously.

The remaining sections of this paper are organized as follows: Sections 2–5 address the fundamentals related to fairness in ML to establish a baseline, covering the relevance of extending the objective of ML, how to define and measure fairness in the context of ML and how to account for biases arising throughout the ML pipeline. Furthermore, this Section 5 deals with the assumed trade-off between accuracy and fairness and presents different approaches that challenge this assumption. In Section 6, the methodology of the analysis is defined, including the selection of data sets, feature pre-processing, and the model and metrics selection to assess fairness and accuracy of outcomes, as well as approaches applied to enhance fairness in ML and their impacts on accuracy. Section 7 describes and evaluates the results and findings for each experiment performed. Finally, this work concludes by presenting the key findings, limitations and further research suggestions.

2. Fairness in Machine Learning

2.1. Relevance of Fairness in Automated Decision Making

Supervised learning models are trained on past data to recognize inlaying patterns and to make future predictions [22]. Based on these models, final decisions are made in sensitive areas of human life, such as whether a person will receive the requested credit, be classified as a repeat offender, be hired for a job, or be diagnosed with a particular disease.

According to the law, those decisions—made automatically by a model based on numbers, or by human judgement—need to be free of discrimination [12]. However, several real world examples emphasize that this requirement is not always met, whether or not sensitive traits are included in the decision making process.

One of the most common known examples is COMPAS, a decision support tool used to estimate the probability of a criminal defendant recommitting a crime by assigning risk scores. Ref. [23] analyzed that the possibility of black defendants being misclassified as risky were more than twice as high compared to white defendants. Therefore, the system invented by Northpointe discriminates on an ethical basis, as race, displaying the protected feature, is not directly included in the analysis. The feature affects the outcome through proxies. Another example is represented by publicly accessible face recognition services developed by Microsoft, Face++ or IBM [24]. Ref. [25] observed that the accuracy for white-skinned males is much greater than for black-skinned females. Consequently, the applied algorithms discriminate not only on an ethical basis, but, additionally, also do so with respect to the gender of an individual.

With AI affecting such sensible areas of human life, and in the context of its evolving future importance, the inclusion of fairness to the ML objective becomes even more important [18]. As [16] points out, the problem needs to be redefined because FML is not

only about "mathematical correctness" and maximizing a model's performance, but also about the integration of "human values" to the ML pipeline to guarantee fair automated decisions. Consequently, to achieve fair decision-making, the objective needs to incorporate both accuracy and fairness.

2.2. Fairness and Group Fairness Measures

There is no universal definition for fairness in ML [17]. According to [18], for example, a decision can be claimed to be fair if it is made in the "absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics".

In general, fairness in ML can be analyzed at the level of the group and of the individual. Group fairness accounts for differences in treatment between groups based on their protected features S [26,27], whereas individual fairness requires each individual to be treated equally, meaning that individuals described by the same features X, including the protected features S, need to be assigned to the same class [28]. In the context of automated decision making, sensible traits, such as gender, age or race [19], are referred to as protected features S.

Table 1 summarizes some popular group fairness metrics based on [17,18,20], including their statistical measures and fairness criteria. All of the metrics are based on the statistical concept of the confusion matrix [17], which are presented in table 2. This work focused on group fairness metrics, Statistical Parity Difference (SPD) and Average Odds Difference (AOD), to assess the disparities between privileged and unprivileged groups based on their protected features [20], such as race or gender [18,19]. Ref. [20] defined three criteria to measure group fairness: independence, separation and sufficiency. Measures built on the criterion of independence ($\hat{y} \perp\!\!\!\perp S$) focus on the predictions made by the model, whereas all other metrics listed rely additionally on the ground-truth labels [17]. All fairness measures are considered to be context-dependent [18] and some impossible to fulfil simultaneously because different measures reflect different perspectives of the stakeholders [16].

Consequently, the selection of fairness measures is of great importance to analyze the relationship between accuracy and fairness.

Table 1. Definitions of Group Fairness (adapted from [17,18,20]).

Measure	Definition	Statistical Measures	Criterion
Statistical Parity [28]	$P(\hat{y}=1\|S=0) = P(\hat{y}=1\|S=1)$	$\hat{y}$	Independence
Conditional Statistical Parity [4]	$P(\hat{y}=1\|S=1,Y=1) = P(\hat{y}=1\|S=0,Y=1)$ $P(\hat{y}=1\|S=1,Y=0) = P(\hat{y}=1\|S=0,Y=0)$	$\hat{y}$	Independence
Equalized Odds [29]	$P(\hat{y}=1\|S=1,Y=1) = P(\hat{y}=1\|S=0,Y=1)$ $P(\hat{y}=1\|S=1,Y=0) = P(\hat{y}=1\|S=0,Y=0)$	$\hat{y}, Y$	Separation
Equalized Opportunity [29]	$P(\hat{y}=1\|S=1,Y=1) = P(\hat{y}=1\|S=0,Y=1)$ $P(\hat{y}=0\|S=1,Y=0) = P(\hat{y}=0\|S=0,Y=0)$	$\hat{y}, Y$	Separation
Predictive Parity [30]	$P(Y=1\|S=1,\hat{y}=1) = P(Y=1\|S=0,\hat{y}=1)$ $P(Y=0\|S=1,\hat{y}=1) = P(Y=0\|S=0,\hat{y}=1)$	$\hat{y}, Y$	Sufficiency

Table 2. Confusion Matrix adapted from [17].

	True Positive	True Negative
Predicted Positive	True Positive (TP) $TPR = \frac{TP}{TP+FN}$	False Positive (FP) $FPR = \frac{FP}{FP+TN}$
Predicted Negative	False Negative (FN) $FNR = \frac{FN}{FN+TP}$	True Negative (TN) $TNR = \frac{TN}{TN+FP}$

Individuals form groups based on their protected features S. In FML those groups are referred to as the privileged ($S = 1$) and unprivileged groups ($S = 0$), with the latter one being the one at disadvantage due to discrimination. To achieve Statistical Parity, the probability of members of the privileged and unprivileged groups being assigned to the positive predicted class needs to be equal [28], implying that the decision made was independent of the group membership with respect to individual traits displayed by the protected feature S. Formally, Statistical Parity of a binary classification problem is summarized by the following equation:

$$P(\hat{y} = 1|S = 0) = P(\hat{y} = 1|S = 1) \tag{1}$$

The definition of Conditional Statistical Parity adds a set of legitimate factors L to the concept of Statistical Parity [4,18]. For example, in the context of credit lending, as described by [17], additional features directly affecting the decision, such as requested loan, age or employment, are included in the analysis. Formally, this indicates:

$$P(\hat{y} = 1|S = 1, L = l) = P(\hat{y} = 1|S = 0, L = l) \tag{2}$$

While the definitions of Statistical Parity fulfil the criterion of Independence, Equalized Odds and Equalized Opportunity apply to the concept of Separation ($\hat{y} \perp\!\!\!\perp S|Y$). Separation assumes the protected feature to be correlated with the target variable Y [20] and the prediction $\hat{y}$ to be conditionally independent of S. On the one hand, Equalized Odds requires equal True Positive Rates (TPR) and equal False Positive Rates (FPR) for both groups [29].

$$\begin{aligned} P(\hat{y} = 1|S = 1, Y = 1) &= P(\hat{y} = 1|S = 0, Y = 1) \\ P(\hat{y} = 1|S = 1, Y = 0) &= P(\hat{y} = 1|S = 0, Y = 0) \end{aligned} \tag{3}$$

On the other hand, Equalized Opportunity is defined by the TP or FN error rate balance, representing the idea that all individuals, independent of their group membership, have an equal possibility of being assigned to the positive predicted class [29]. Hence, members of the privileged and unprivileged groups are defined as have matching TPR or False Negative Rates (FNR).

$$\begin{aligned} P(\hat{y} = 1|S = 1, Y = 1) &= P(\hat{y} = 1|S = 0, Y = 1) \\ P(\hat{y} = 0|S = 1, Y = 0) &= P(\hat{y} = 0|S = 0, Y = 0) \end{aligned} \tag{4}$$

The definition of Predictive Parity fulfils the third fairness criterion, namely Sufficiency ($Y \perp\!\!\!\perp S|\hat{y}$). To achieve Sufficiency the protected feature S is assumed to be conditionally independent of the true class Y given the predicted class assignment $\hat{y}$. The idea behind it indicates that the protected feature is covered by the prediction of the target variable [20]. Therefore, Predictive Parity is fulfilled if both groups have equal positive predicted values. This implies that the positive predictions should be equal for all values of the protected feature [30].

$$\begin{aligned} P(Y = 1|S = 1, \hat{y} = 1) &= P(Y = 1|S = 0, \hat{y} = 1) \\ P(Y = 0|S = 1, \hat{y} = 1) &= P(Y = 0|S = 0, \hat{y} = 1) \end{aligned} \tag{5}$$

2.3. Individual Fairness

Individual fairness considers each individual independently, instead of assuming a decision to be fair on a group level. Fairness Through Awareness makes use of distances to estimate the similarity between individuals. The main idea is that similar individuals are treated equivalently based on their sets of features X, including the protected features S [28]. In opposition, Fairness Through Unawareness does not include the protected features S in the analysis. Hence, individuals with the same set of features X are assigned to the same target class [31]. However, the effect of the protected feature might be captured by other features from the considered feature set [23]. Moreover, causal analysis is used to explore the relationship between the whole data set, including X, S and Y. According to [32] "a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic

group"—implying Counterfactual Fairness. Therefore, in the sense of causal graphs the predicted class needs to be independent of a predecessor of the protected feature [17]. An overview of the above-mentioned individual fairness measures can be found in Table 3.

Table 3. Definitions of Individual Fairness adapted from [17,18,20].

Measure	Paper	Concept
Fairness Through Awareness	[28]	Similarity
Fairness Through Unawareness	[31,32]	Similarity
Counterfactual Fairness	[32]	Causality

3. Sources of Bias in the Machine Learning Pipeline

Discrimination means putting certain individuals or groups of individuals at a disadvantage based on their sensitive traits [18]. In the context of ML, this implies that models are biased in the way that certain individuals are assigned to a negative class ($\hat{y} = 0$) based on their sensitive features S, such as gender, age or race [19]. The protected features serve as an input for the model, which represents a "collection of decision rules" [2], to classify an instance as good or bad, creditworthy or not, or to assess the probability of a criminal relapsing. For binary classification problems, discrimination comes with a polar target variable. The predictions are polar in the sense that one outcome is more beneficial compared to the other outcome [2].

Discrimination is often associated with biased data. This assumption treats data like a static and independent entity and discards the underlying procedures. Data collection is a continuous process reflecting historical patterns, and human norms and judgements, as well as the purpose of usage [33].

3.1. Sources of Bias in the Data Collection Phase

Each stage of the ML pipeline, including data collection, model development, evaluation or even model deployment, involves decisions [33] which lead to final outcomes, and, therefore, could be a potential source of bias. There has been a great amount of research in recent years on developing strategies to enhance fairness in outcomes [21,34,35]. Ref. [20] categorized the categories as data pre-processing, model in-processing and outcome post-processing. In [7] it is claimed that the first stage, namely data collection, is the origin of the trade-off. Therefore, this work concentrates on the data collection process to explore the relationship between accuracy and fairness. A basic principle of ML states: garbage in, garbage out [36,37]—which relates to bad quality data leading to undesirable outcomes. This stage involves sampling the group of interest and selecting and measuring characteristics and labels [33]. Representation bias, or sampling bias, is based on the definition of the target population. Even if the data is perfectly measured, certain individuals or groups may be unintentionally disadvantaged because they are underrepresented, whereas measurement bias refers to the selection of characteristics and their measurement [33]. In addition, characteristics within the selected group of characteristics may be interrelated and hide their true effects on the outcome [38]. Furthermore, historical bias occurs when data from the past are used to make predictions about the present and future [3,33]. Finally, decisions that affect a person's life, such as attitudes, health care, or lending, are based on human judgment and, therefore, may not correspond to the ground truth, although ML models assume that a data set's labels reflect the ground truth [2,3,38].

3.2. Model Development

The model development summarizes the model training and leads to the final decision. Bias can be introduced by using a "one-size-fits-all model" referred to as aggregation bias [33]. Although the data set might contain different groups, the model does not depict these variations in the underlying conditional distributions and, hence, leads to inappropriate predictions [39]. Moreover, during the training process hyperparameters are chosen to maximize a certain objective. However, the defined objective might not be

suitable for the problem [33], so that, for fairness in ML, optimizing the model with respect to accuracy is insufficient. In addition, algorithmic processing bias deals with biases that emerge when accounting for fairness. By implementing mitigation strategies to account for fairness during model development, bias is again introduced in the sense that the model is not objective but, instead, tries to account for differences between groups or individuals with respect to their sensitive features S [40].

3.3. Model Evaluation

Furthermore, bias is introduced assessing model performance. One source is presented by the choice of metrics used to evaluate a model. The context of application needs to be considered critically, since a model achieving great accuracy does not necessarily indicate that the model performs well in reality [33]. For example, in the context of healthcare, FN come at greater costs than FP. Additionally, evaluation bias arises due to benchmarking, because the data used to measure performance does not display the target group [39].

3.4. Model Deployment

The last step of the ML pipeline is defined by model deployment. Deployment bias describes the problem occurring if the purpose of the model does not match its application [33]. The "framing trap", defined by [41], captures this problem of mismatch, wherein a model trained for risk assessment might as well be used to predict if an individual will show up in court [33]. Moreover, the outputs produced by the model need to be interpreted adequately. Interpretation bias emerges because predictions are built on probabilities performed by the model [42]. For the example of COMPAS, the court is prepared with risk assessment scores [23] which they need to interpret by themselves and, consequently, user judgement is introduced. Non-transparency of outcomes refers to the "black-box problem" described in Section 2.1, whereby the rules of a ML model leading to a prediction are not transparent for the user and, therefore, might be incomprehensible [42]. Automation bias supports the assumption that automated decision making magnifies discrimination [3]. As already indicated by the non-transparency of outcomes bias, the user does not know the underlying assumptions and does not know about potential hidden biases. Hence, the user assumes that the model makes the right decision. As a result, the user does not critically analyze the predictions made by the model [42].

4. Bias Mitigation Strategies

Strategies to account for discrimination in automated decision making can be classified along the pipeline as follows: (1) pre-processing or data-based mitigation; (2) in-processing or model-based mitigation; (3) post-processing mitigation methods [20]. Pre-processing mitigation strategies correspond to the stage of data collection in the ML pipeline and are used to ensure that the predefined protected feature does not impact the outcome negatively by modifying the feature space [20]. The biases inherent in the data itself are removed before model training to account for fair outcomes [22]. Bias mitigation at the stage of model development includes diverse approaches. In-processing mitigation strategies are modifications of model development, e.g., changing the loss function, adjusting the learning algorithm itself or compositional and adversarial approaches [22]. However, since these approaches are only suitable for certain optimization problems [20], in-processing mitigation strategies come with less generality. Post-processing mitigation approaches occur at the stage of model evaluation. They are applied on potentially biased, trained classifiers to ensure fairness with respect to the sensitive traits of an individual [22]. Post-processing is independent of the modeling process and, hence, applies to any black-box classifier. Additionally, the models do not need to be trained again, since post-processing mitigation only affects the outcomes and represents the only possibility to control for fairness if access to the ML pipeline is not given [20].

In this work, we focused on the Reweighing (RW) approach [21], one of the most popular pre-processing mitigation strategies. With Reweighing (RW), each sample of the

data set, representing an individual, is assigned a weight based on their ground-truth label Y and predefined protected feature S. The main idea constitutes re-balancing the data set with respect to the protected feature to ensure fairness in outcomes [21].

4.1. Pre-Processing Strategies

Pre-processing strategies correspond to the stage of data collection in the ML pipeline and are used to ensure that the predefined protected feature does not impact the outcome negatively by modifying the feature space [20]. The biases inherent in the data itself are removed before model training to account for fair outcomes [22].

Probably the most popular pre-processing mitigation strategy is Reweighing (RW). The technique was invented by [21] and, instead of altering the data by changing the ground-truth labels, as done for Massaging [43], each sample of the data set, representing an individual, is assigned a weight based on the ground-truth label Y and predefined protected feature S. The main idea constitutes re-balancing the data set with respect to the protected feature to ensure fairness in outcomes [21].

Instead of re-sampling or re-labeling the data, representation learning is applied to mitigate biases arising at the beginning of the ML pipeline [22]. Optimized Pre-processing makes use of a probabilistic framework to generate a randomized mapping of the training and test sets that controls for discrimination. To conclude, the data set $D = (X_i, S_i, Y_i)$ is transformed into a new data set $\tilde{D} = (S_i, \hat{X}_i, \hat{Y}_i)$. The randomized mapping needs to fulfil the following properties: First, the effect of the protected feature on the transformed mapping of the outcome is limited; Second, large changes are suppressed by distortion constraints; Third, statistical distribution of the mapping needs to be close to the real distribution to preserve utility [44]. Another approach falling into this category is the so-called Disparate Impact Remover. In contrast to Optimized Pre-processing, it affects only the feature set X, instead of X and the ground-truth label Y, leading to a new data set $\tilde{D} = (S_i, \hat{X}_i, Y_i)$ [45].

The approach of Learning Fair Representations [34] is built on the work by [28], who explored probabilistic mapping that fulfils individual and group fairness simultaneously. The main idea of [34] is to achieve a set of intermediate representations that fulfil two criteria: first, the representation reflects the acquired data as well as possible and, second, the representation is not able to assess if an instance originally belonged to the protected group. To summarize, the mapping is designed to keep all information, excepting group affiliation, based on the protected feature. In addition, the algorithm comes with hyperparameters to optimize outcomes with respect to fairness and accuracy [34].

The methods described cover only a subsection of pre-processing mitigation strategies to improve fairness in ML. Further relevant techniques include a Variational Fair Autoencoder [46], Rule Protection [47], Adversarial Learned Fair Representations [48] and others, summarized by [22].

The analytical part of this work focused on pre-processing mitigation strategies, because label bias, in particular, plays a critical role when it comes to model performance, which is often avoided in research [3]. Moreover, bias arising at the stage of data collection affects the entire ML pipeline and by correcting for bias in the first stage flexibility for further steps is maintained [20,44].

4.2. In-Processing and Post-Processing Strategies

Nevertheless, to provide a general overview and understanding of bias mitigation throughout the ML pipeline, in-processing and post-processing are described broadly in the following. Moreover, the advantages and disadvantages of each strategy are discussed.

In-processing mitigation strategies are modifications of model development, which represent the second step of the ML pipeline, as illustrated by Figure 1. While one of the advantages of pre-processing is retaining flexibility along the ML pipeline, this illustrates a disadvantage of in-processing strategies. At the same time, modifying the model leads to less generality, since those approaches are only suitable for certain optimization prob-

lems [20]. Bias mitigation at the stage of model development includes diverse approaches, such as changing the loss function, adjusting the learning algorithm itself or compositional and adversarial approaches [22].

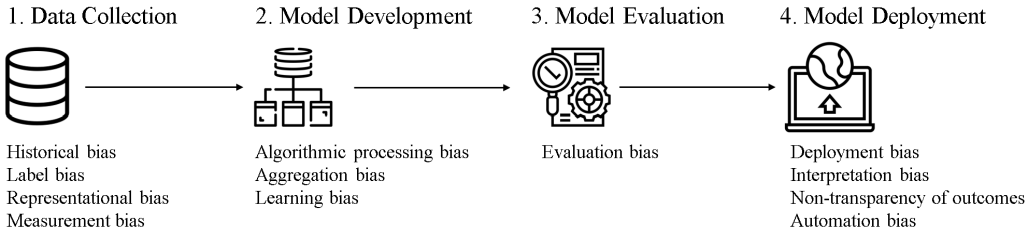


Figure 1. Potential biases of the ML pipeline.

Approaches falling into the category of post-processing mitigation occur at the stage of model evaluation. They are applied on a potentially biased, trained classifier to ensure fairness with respect to the sensitive traits of an individual [22]. On the one hand, adapting the outcomes might lead to a reduction of utility. On the other hand, post-processing is independent of the modeling process itself and hence applies to any black-box classifier. Additionally, the models do not need to be trained again, since post-processing mitigation only affects the outcomes and it represents the only possibility to control for fairness if access to the ML pipeline is not given [20].

5. Accuracy and Fairness

In ML, accuracy is defined as the objective of the classification task [16]. Accuracy metrics are applied to measure model performance by comparing the "label alignment" of the predictions made by the model, trained on a subset of the data set, to the ground-truth labels of the test set [3]. Due to label bias, the concept of accuracy in FML is challenged. Although label bias is a common and known problem, only a few works deal with this problem of discrimination when evaluating model performance based on the assumed ground-truth labels [13]. As explained in Section 3.1, the ground-truth labels might be discriminating. If the ground-truth labels are wrong, this, consequently, leads to inaccurate decisions [3]. Based on the assumption that the ground-truth labels are fair, predictions become unfair, as emphasized by the examples given in Section 2.1. Since automated decision making influences human life, fairness needs to be granted to avoid systematic discrimination.

Fairness metrics measure the difference in decisions between groups or individuals based on the protected features included in the modeling process [3]. To achieve group fairness, the decision made, illustrated by the prediction $\hat{y}$, needs to be independent of the sensitive feature S, summarized by the Independence criterion [20] representing the "we're all equal worldview" [7]. The sensitive traits of an individual do not reflect his or her performance capability. Consequently, the individual's protected features should not have an effect an the assigned class labels [13]. Therefore, the objective function should not only focus on the performance of an algorithm measured by accuracy metrics, but put fair accuracy in the center of the objective.

5.1. The Trade-Off Assumption between Accuracy and Fairness

The great majority of research on the topic of FML assumes a trade-off between accuracy and fairness [3–7]. It is claimed that enhancing fairness in automated decision making leads to an unavoidable loss of accuracy [3], also referred to as the "cost of fairness" [4]. The authors in Ref. [7] argued that the data collection stage is the source of the assumed trade-off. Consequently, their analysis of the relationship between accuracy and fairness concentrated on pre-processing strategies. The authors in [5] state that accuracy and fairness can only be maximized simultaneously if the target variable is fully independent of the protected feature. Yet, due to noisier mappings, this is not achievable in reality [13]. Pre-processing techniques are used to account for noisier mappings. If only the training

set is cleaned for label bias, the performance of the model is still evaluated on the potentially discriminating ground-truth labels, which comes at the cost of accuracy and leads to biased measurements [13]. The study in [49], in the specific application area of credit scoring, investigates the profit implications of fairness and also discusses the assumed trade-off between fairness and accuracy in that particular domain. Our study investigated two different application domains and also further analyzed the trade-off from different perspectives and well-known ML practices.

5.2. State-of-the-Art Approaches

In contrast to the trade-off assumption, some researchers argue that a negative relationship is often taken for granted without being verified and highlight that the assumed trade-off is not necessarily a given [3,8,10–14].

Although fairness and accuracy are assumed to be in opposition to each other, [12] explains that maximizing accuracy simultaneously enhances fairness, stating that increasing accuracy leads to a decrease in errors produced by the model, represented by False Negatives (FN) and False Positives (FP), and, consequently, has a positive effect on the fairness in outcomes. Moreover, collecting more features to improve the description of an individual is claimed to increase fairness and accuracy, referred to by the term active fairness [50]. Ref. [11] analyzed the effect of feature set and training set size on fairness. Their approach built on two principles: feature set size defining the "knowledge level" and training set size defining the level of experience, both of which are commonly known to enhance accuracy of predictions. They found that enlarging the number of features used for model training improves fairness, whereas increasing the size of the training set has no significant effect, or even a negative impact, on fairness in outcomes. Their explanation is that, by introducing more biased data points, the model is able to depict those patterns more accurately which, consequently, leads to greater unfairness. [10] augment the most realistic points of a synthetic data set to the original data set to account for the problem of label bias. The findings emphasized that fairness can be enhanced without causing a significant loss of accuracy and that increasing one fairness metric simultaneously increases other metrics. This approach does not rely on mitigation strategies to enhance fairness, which decreases its complexity and also leads to greater transparency of the modeling process. Other approaches to account for the assumed trade-off between fairness and accuracy include post-processing strategies, such as the confidence-based approach proposed by [14]. They consider the problem of label bias and explore a way to optimize the trade-off by using shifted decision boundaries, based on the theory of margins. The method does not negate the assumed trade-off, but makes it accessible by regularizing the decisions made by the model. The labels corresponding to small signed confidences are reversed. Based on the theory of margins, flipping labels does not have a major impact on accuracy. Their approach can be applied to any ML algorithm which outputs display confidence measures, such as boosting or logistic regression. The authors in [15] make use of group-specific thresholds to adjust the TPR among individuals. Their results emphasize that fairness can be achieved by a minor decrease in accuracy.

6. Data and Methodology

The following experiments were carried out:

1. **Optimized modeling: Does maximizing accuracy simultaneously lead to an increase in fairness?** Enhancing accuracy with respect to ACC and BACC is expected to decrease the corresponding error rate and, consequently, to simultaneously increase fairness [12].
2. **Discrimination threshold: Does adapting the threshold of classification in the interest of maximizing accuracy improve fairness?** The LRC is optimized with respect to the F1 score. Adapting the threshold of classification is expected to increase accuracy. Furthermore, balancing the samples representing the privileged and unprivileged groups might simultaneously increase fairness.

3. **Bias mitigation: How does the objective of maximizing fairness impact accuracy?** To analyze the effect of bias mitigation on accuracy and fairness, the popular mitigation strategy RW was performed. According to the research, assuming the trade-off between accuracy and fairness to be unavoidable, accuracy is expected to decrease if fairness is maximized.
4. **Feature set size: Does the number of features used to make a decision influence accuracy and fairness?** Increasing the feature set size, corresponding to greater knowledge being available about an individual, is expected to increase accuracy and fairness [11].
5. **Training set size: Does the quantity of available training data influence accuracy and fairness?** Increasing the training set size is expected to increase accuracy but to also negatively impact fairness, according to [11].
6. **Data augmentation: How do accuracy and fairness change if synthetic data points are augmented to the data set?** The addition of realistic, synthetic data points to account for different labels with respect to the unprivileged and privileged groups is assumed to increase fairness and accuracy simultaneously [10].

Our work replicates the data simulation approach, based on [10] and the impact of training set and feature set size, as proposed by [11], as well as the pre-processing mitigation strategy, namely RW [21]. The same fairness and accuracy measures are applied to enhance comparability across the applied approaches used to increase fairness in ML and to explore the relationship between fairness and accuracy.

6.1. Data Set Selection

The following analyses were based on two datasets: COMPAS [23] and LAW [51]. The prediction task for COMPAS is to assess whether an individual will become recidivist within two years after imprisonment. For LAW, an individual is classified as to whether he or she will pass the bar exam or fail the bar exam. COMPAS is a system invented by Northpointe to assess the risk of recidivism and the data set was published in 2016 by *ProRepublica*. This data set contains a total of 52 features. However, only the variables listed in Table A1 in Appendix A were used for further analysis. Figure 2 illustrates the distribution of individuals assigned to be re-offenders within the considered time period per group. The pie charts in Figure 2 emphasize that African-American individuals predicted to become recidivist made up more than half of the total group and exceeded the share obtained for Caucasians.

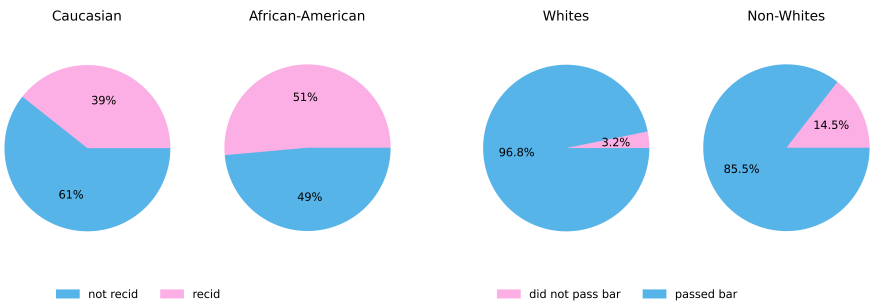


Figure 2. Label distribution according to protected feature. (**a**) COMPAS; (**b**) LAW.

The other data set, LAW (http://www.seaphe.org/databases.php, accessed on 12 May 2022), is an educational data set used in FML. This data was collected via surveying more than 20,000 law students from 167 different law schools in the United States in 1991 [51]. Figure 2 illustrates that the majority of students passed the bar exam. However, the share of non-white students ($S = 0$) failing the bar exam was more than four times greater compared to white students ($S = 1$). The group of non-white students summarized

individuals belonging to race groups described as Black, Hispanic, Asian and others. The total ratio of Whites compared to Non-Whites amounted to 17,524 (84.10%): 3313 (15.89%) summarizing a total of 20,836 observations. Consequently, for this data set (in contradiction to COMPAS) the unprivileged group was underrepresented. Moreover, the ratio of the positive and negative target class was imbalanced: 17,524 (84.10%): 3313 (15.9%). Although the amount of samples belonging to the negative class ($Y = 0$), representing failure of the bar exam, was minor compared to positive samples, the majority of negative class labels corresponded to the unprivileged group.

6.2. Data Pre-Processing

For comparability, both data sets were pre-processed equivalently. As for encoding of variables, for LAW the favorable label indicated passing the bar exam, while for COMPAS the admired outcome was not being classified as recidivist. To ensure uniform encoding, the positive outcome corresponded to $Y = 1$, whereas the negative class was defined by $Y = 0$. For LAW, the favorable label indicates passing the bar exam. Equivalently, the unprivileged group is identified by $S = 0$ and the privileged group is represented by $S = 1$. For COMPAS, only individuals belonging to the African-American and Caucasian races were kept and binary encoded, since both the considered groups made up the majority of the data set. As the majority of data points of LAW belonged to white individuals, all other variations of the protected feature race were summarized by Non-Whites. Additionally, all categorical features were one-hot encoded.

Next, all rows containing missing values were dropped. Additionally, the features were selected based on the *AIF360* open-source toolkit [35] for COMPAS and on Google Research (https://github.com/google-research/google-research/blob/master/group_agnostic_fairness/data_utils/CreateLawSchoolDatasetFiles.ipynb, accessed on 12 May 2022) for LAW:

COMPAS: {*race*, *sex*, *age*, *juv_felt_count (Juvenile felony count)*, *juv_mids_count (Juvenile misdemeanor count)*, *juv_other_count (Juvenile other defenses count)*, *priors_count (Prior offenses count)*, *c_charge_degree (Charge degree of original crime)*, *two_year_recid (Rearrested within two years)*}; LAW: {*race*, *gender*, *lsat (The student's LSAT score)*, *ugpa (The student's undergraduate GPA)*, *zfygpa (The first year law school GPA)*, *DOB_yr (Year of birth)*, *zgpa_felt_count (The cumulative law school GPA)*, *family_income (The student's family income bracket)*, *part_time (Working full- or part-time)*, *cluster_tier (Tier of school)*, *weighted_lsat_ugpa (Weighted LSAT and UGPA score)*, *pass_bar (Passed bar exam on first try)*}.

To use the fairness measures provided by the *AIF360* toolkit, the data needed to be transformed in a certain way. The target feature set y not only included the ground-truth labels, but additionally the protected feature. For both feature sets X and y, the protected feature was set as index, to enable the use of fairness metrics and mitigation strategies provided by the toolkit. After setting the protected feature as index, it was dropped for the target set y and additionally maintained for X. Finally, some ML algorithms required scaled data to achieve reasonable results. The $StandardScaler()$ function was applied on the training sets before training and test sets were transformed accordingly.

6.3. Model Selection

To account for the black-box problem, the basic classifier LRC was applied to the considered data sets. Based on the collected features the LRC estimated the probability of an individual belonging to the positive or negative class [52]. RFC combines multiple single decision trees. The classifier is expected to increase accuracy due to averaging over these multiple single learners and, consequently, decreasing variability [53]. The GBC is applied to optimize the differentiable loss functions of the single trees to "boost" accuracy [54]. For the following analysis only the performance results for the LRC were documented, due to its popularity and the low calculation costs compared to the other two classifiers [52]. The classifier was optimized to maximize accuracy. Hence, a grid search was performed to select the best hyperparameters with respect to ACC and BACC. Due to data imbalance

and the validation set only reflecting a small subset of the whole data set, the optimal hyperparameters were evaluated on repeated stratified 5-fold Cross Validation (CV).

Moreover, due to the nature of a binary classification task, the classification threshold τ plays an important role in model evaluation—especially in the presence of imbalanced data sets. By default, the threshold was set to $\tau = 0.5$. However, this might not always reflect the best interpretation of the estimated probabilities. The discrimination function prepared by Yellowbrick (https://github.com/DistrictDataLabs/yellowbrick/blob/main/docs/api/classifier/threshold.rst, accessed on 12 May 2022) was used to determine the optimal thresholds for COMPAS and LAW. In the interests of finding the optimal cut-off threshold τ to maximize accuracy measured by errors, while at the same time accounting for stability of performance and data imbalance, the provided function was carried out using stratified 10-fold CV.

6.4. Fairness Measures

The choice of metrics to evaluate model performance with respect to fairness is a critical decision in model evaluation, because different fairness metrics lead to different results [35]. To evaluate fairness with respect to the disparate representations, SPD was applied. SPD is defined by the following formula [35]:

$$SPD = P(\hat{y} = 1|S = b) - P(\hat{y} = 1|S = a) \quad (6)$$

Additionally, AOD was used to analyze fairness, based on the disparate errors of the system, to achieve Sufficiency. AOD was estimated based on Equation (7).

$$AOD = [FPR(S = 0) - FPR(S = 1)] + [TPR(S = 0) - TPR(S = 1)] \quad (7)$$

For both metrics, a value closer to zero indicated more fairness and represented less bias. If the measurements displayed a negative value, this indicated that the privileged group was at an advantage, whereas, for a value greater than zero, the unprivileged group was benefiting [35].

6.5. Accuracy Measures

To assess the correctness of predictions made by the trained classifier, and to consider disparate errors, the following metrics were applied. First, Accuracy (ACC) measured the correctness of the predictions compared to the ground-truth labels. The ground-truth labels were assumed to be reliable. Second, Balanced Accuracy (BACC) was applied to account for the imbalance of the LAW data set. BACC was defined by the average of the sum of sensitivity (TPR) and specificity (TNR). If the labels of a data set were balanced, the results achieved by ACC and BACC should be similar. The F1 score summarizes both perspectives and, consequently, is often referred to as the "harmonic mean between precision and recall" [55].

6.6. Performance Stability

Ref. [56] showed that performing a single train–test split to evaluate a model's performance is insufficient, due to the instability of model performance. According to their findings, the performance evaluated on fairness metrics, in particular, appeared to vary a lot for various train–test splits. They recommend using a moderate number of train–test splits to account for performance instability. In our study, instead of randomly splitting the data into training and test sets several times, for k-fold CV the data was split once into ten folds $(k = 10)$ with one fold representing a hold-out set, referred to as the test set to evaluate the model's performance. K-fold CV ensures that each available data point is used for training and testing, while for repeated k-fold CV this is not necessarily so, especially in the case of data imbalance. Due to the given target class imbalance for the considered data sets, stratified 10-fold CV was applied. Stratifying samples ensures that each fold is representative and reflects the underlying distribution of the considered data set [57].

6.7. Reweighing

The objective of this stage was to maximize fairness. To analyze the impact of RW on fairness and accuracy, the analysis was divided into two stages. First, based on the *scikit-learn* compatible API reference of *AIF360* (https://aif360.readthedocs.io/en/stable/, accessed on 12 May 2022), each sample was rebalanced according to its protected feature and target class, producing four different weights for COMPAS and LAW. The compatible version did not enable altering of the test set. Consequently, only the sample weights of the training set were transformed and evaluated applying stratified 10-fold CV. The performance was evaluated on all predefined accuracy and fairness metrics to enable comparison to the baseline results.

However, [10] recommend reporting the achieved performance results for an altered training set evaluated on a hold-out test set, as well as on an entirely processed data set, including adapted training and test data. Therefore, for the second stage, the RW function prepared by the old API was used to pre-process the training and test sets. Since the transformed sample weights were lost, if the transformed data sets were converted back to a *pandas* data frames and the old API was not compatible with common meta-programming tools provided by *scikit-learn*, such as CV [58], the data was split ten times into training and test sets to account for stability during performance evaluation on the rebalanced training and test sets. The results were compared to baseline results of a LRC assessed by equivalent splits.

6.8. Feature and Training Set Size

A general practice in ML to maximize accuracy is to increase the feature set size or to expand the training data used for a model's development [59]. Ref. [11] analyzed the effect of feature and training set sizes on fairness. They claim that feature set size has a positive impact on fairness, while increasing the size of the training set leads to the opposite effect. Furthermore, for the authors balancing the data with respect to the members belonging to the unprivileged or privileged group or by applying bias mitigation strategies, means accuracy and fairness do not need to be conflicting regarding enlarged training data.

In our study, first, the feature set size was analyzed by performing stratified 10-fold CV on sets of augmented features and averaging the performance of the LRC. The protected feature needs to be included in the feature set at any time, to enable the use of fairness measures. For both data sets the smallest considered feature set size amounted to three features, including the protected feature.

Second, the impact of enlarged training data on fairness was studied. For this purpose, the data sets were split into training in test sets, with the training sets ranging in size between 10–90%. At least ten percent of the data were kept for model training or evaluation. The average impact on accuracy and fairness measured by the corresponding metrics was reported on ten random train–test splits. Furthermore, based on the recommendations of [11], the same procedure was applied after reweighing the training data to account for inherent biases.

6.9. Data Augmentation

The approach of data augmentation used in this work was based on [10]. By deriving synthetic data and adding the most realistic synthetic data points to the original data sets, the authors examined whether fairness could be enhanced by holding accuracy constant.

As part of the data augmentation, the synthetic dataset was first generated as a copy of the original dataset with reversed values for the protected feature *race*. Individuals belonging to the privileged group were defined by $S = 0$ and the unprivileged group by $S = 1$. All other features X and the labels Y were equivalent to the original data set. Second, K-means was used to estimate the clusters of the original data sets, based on their protected features and labels. Consequently, "a set of cluster centers are defined for every protected

feature value and every label value" [10], described by S and Y. Third, the synthetic data points were sorted according to their realism scores, (8) [10] as defined by the formula:

$$RealismScore = \frac{1}{max\{d(c_1, p)\}, max\{d(c_2, p)\}, ..., max\{d(c_j, p)\}} \quad (8)$$

Here, we estimated the distance between each synthetic data point p and the predefined cluster centers $c = \{c_1, c_2, ..., c_k\}$ based on the original data set. Fourth, after sorting the synthetic data set the j most realistic data points, with j ranging from 0–100 percent, were augmented to the original data set. Finally, the performance was measured in two different ways. First, accuracy and fairness were evaluated on test sets, including augmented data. Second, the test sets were unaltered to preserve the real world distribution, while the model was trained on augmented training sets.

7. Results

7.1. Baseline Results

The first step of the analysis deals with basic classifiers to provide baseline performance results, which are compared in the following analysis to studied approaches to explore the relationship between accuracy and fairness. The results displayed in Table 4 illustrate the average performance results of the LRC reported for stratified 10-fold CV.

Table 4. LRC baseline results. Performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness.

	ACC	BACC	F1	SPD	AOD
COMPAS	0.675	0.669	0.713	−0.261	−0.230
LAW	0.952	0.566	0.975	−0.061	−0.125

Whereas ACC and BACC were quite similar for COMPAS, the accuracy measures differed a lot regarding the LAW data set, due to significant class imbalance. The BACC value of 0.566 indicates that the LRC was not able to accurately predict outcomes with respect to the imbalanced target classes. Furthermore, SPD and AOD show that the predictions made for COMPAS were biased and, therefore, unfair. The negative value for SPD indicates that the probability for the privileged group to be assigned to the positive class on average exceeded the probability estimated for the unprivileged group. For LAW the results for SPD and AOD differed in their magnitudes. While the model appeared to be quite fair regarding the probabilities of Whites and Non-Whites being assigned to the positive class, the AOD value evidenced greater unfairness. Whereas SPD measured the equality of distributions, AOD was calculated based on the difference of the FPR and the TPR for the privileged and unprivileged groups. Therefore, the data imbalance of the LAW data set appeared to negatively influence fairness in outcomes. Unfairness in the case of LAW seemed to be due to representational bias of the data at hand. These discrepancies highlight the importance of the definition and selection of appropriate fairness metrics with respect to the problem formulation.

In addition, Figure 3 displays the results for RFC and GBC representing the predefined accuracy measure ACC on the y-axis and fairness on the x-axis. Each cross applies to one of the classifiers, illustrating the variations in accuracy and fairness. The horizontal lines mark the standard deviations for the fairness metrics and the vertical lines refer to the standard deviations achieved for ACC. The centers of the intersections, represented by dots, define the mean performance values for stratified 10-fold CV. To ensure comparability, the axes for LAW and COMPAS were set to the same sizes.

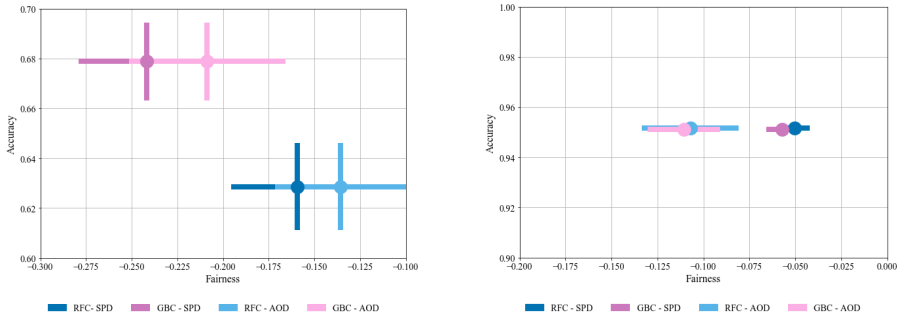


Figure 3. Stability of fairness and accuracy for RFC and GBC. Performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. (**a**) COMPAS; (**b**) LAW.

Accuracy and fairness measured by ACC and SPD appeared to be more stable for the LAW data set compared to COMPAS, which could be due to more accurate results with an ACC score of 0.95. For both data sets, the results achieved by SPD and AOD differed for variations in training and test data, as illustrated by the standard deviations. These findings are in line with those of [56]. Fairness measures are more sensitive regarding variations in the input data, compared to measures assessing the correctness of predictions.

Although the pre-processing and modeling steps were equivalent for both data sets, the results differed a lot, which is emphasized by the use of colors in Figure 3. The results achieved by the RFC are illustrated by blues, whereas the outcomes for GBC are displayed by shades of pink. Moreover, the bright colors represent the results for AOD. For COMPAS both predefined fairness metrics performed quite similarly with respect to the considered classifier. However, with respect to accuracy, the GBC outperformed the RFC, while the latter achieved better outcomes with respect to fairness. For the LAW data set the results created a different picture. The difference of performance was based on the fairness measures rather than the applied classifiers. Additionally, SPD performance results seemed to be more stable, compared to AOD, as indicated by smaller standard deviations.

7.2. Optimized Modeling

The second stage deals with the optimization of the LRC using a grid search. The objective of this stage is to maximize accuracy measured by ACC and BACC. As [12] states, maximizing model performance should simultaneously lead to an increase in fairness. The best hyperparameters determined by a grid search for the LRC were defined as follows: COMPAS: $\{C = 0.1, class\_weight = none, penalty = l1, solver = saga\}$ to maximize ACC and BACC, LAW: $\{C = 0.1, class\_weight = none, penalty = l1, solver = saga\}$ to maximize ACC, LAW: $\{C = 0.01, class\_weight = balanced, penalty = l1, solver = liblinear\}$ to maximize BACC. Table 5 illustrates LRC results for tuned hyperparameters.

Table 5. LRC results for tuned hyperparameters. Performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. The percentages in brackets display the differences compared to the baseline results, as presented by Table 4.

	Objective	**ACC**	**BACC**	**F1**	**SPD**	**AOD**
COMPAS	ACC + BACC	0.675 (=)	0.669 (=)	0.714 (+0.14%)	−0.259 (+0.77%)	−0.228 (+0.87%)
LAW	ACC	0.952 (=)	0.562 (−0.71%)	0.975 (=)	−0.056 (+8.2%5)	−0.117 (+6.4%)
	BACC	0.793 (−16.7%)	0.813 (+43.64%)	0.879 (−9.85%)	−0.410 (−572.13%)	−0.288 (−130.4%)

For COMPAS, the hyperparameters to maximize ACC and BACC were equivalent. In this case, maximizing accuracy slightly improved fairness. Due to data imbalance for LAW the optimal hyperparameters differed with respect to the considered metrics of

accuracy. These results illustrate that the interpretation of accuracy was as equally crucial as the definition of fairness. Whereas maximizing for ACC increased fairness by 6–8%, optimizing the classifier with respect to BACC led to a major loss of fairness. In particular, SPD decreased significantly. Instead of improving fairness in outcomes, it led to even greater unfairness due to the enlarged disparities between the unprivileged and privileged groups. The difference between the two groups increased due to disregarding the protected feature *race*, which was assumed to have a significant impact on class assignment.

To conclude, for balanced data, maximizing ACC or BACC potentially increases fairness. For imbalanced data sets, maximizing BACC might have a negative impact on fairness, since maximizing model performance with respect to BACC does not account for the impact of the protected attribute on the target class. Therefore, the choice of accuracy and fairness metrics plays a key role in analyzing their relationship.

7.3. Discrimination Threshold

A further step is to maximize the F1 score by adapting the discrimination threshold τ of classification. Balancing precision and recall is expected to increase fairness by prioritizing errors, that comes at the cost of fairness. Table 6 shows the LRC performance results for adapted discrimination thresholds.

Table 6. LRC performance results for adapted discrimination thresholds. Performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. The percentages in brackets display the differences compared to the baseline results as presented by Table 4.

	τ	ACC	BACC	F1	SPD	AOD
COMPAS	0.42	0.653 (−3.26%)	0.636 (−4.93%)	0.732 (+2.66%)	−0.180 (+31.03%)	−0.156 (+32.17%)
LAW	0.45	0.952 (=)	0.551 (−2.65%)	0.975 (=)	−0.042 (+31.15%)	−0.092 (+26.4%)

Compared to the results for basic modeling without tuning, adapting the classification threshold τ accordingly led to better results regarding fairness. For both data sets, adjusting the threshold of discrimination down led to an increase of fairness by 26–31% compared to the baseline results. Setting the threshold lower shifted the cut in favor of the unprivileged group. Therefore, adapting the classification threshold τ was identified to have a positive impact on fairness.

How does the discrimination threshold τ impact accuracy? The F1 score was enhanced, because the models were optimized with respect to the considered score. However, ACC and BACC appeared to slightly decrease. The effect for COMPAS was greater, compared to LAW, due to the already accurate baseline results achieved for LAW. The only metric affected for LAW was BACC. Although modifying the threshold τ is known to be a suitable approach to account for data imbalance, BACC decreased for LAW and COMPAS compared to the baseline performance results. These findings indicate that adjusting the discrimination threshold causes a loss of TPR and TNR, while enhancing fairness in outcomes.

To summarize this stage of the analysis, adapting the discrimination threshold of classification appears to be a suitable approach to maximize accuracy, while at the same time enhancing fair outcomes. Maximizing accuracy measured by the F1 score leads to an increase of fairness. Additionally, accuracy measured by ACC is not affected, or only slightly decreases. Therefore, this stage already emphasizes that the relationship between accuracy and fairness does not necessarily need to be a trade-off and, moreover, underlines the importance of the selected metrics in evaluating performance.

7.4. Reweighing

The results displayed in Table 7 report the average results for the LRC on reweighed training data with the test sets remaining unaltered. Compared to the baseline results achieved by the LRC, fairness increased for both LAW and COMPAS and accuracy de-

creased, with one exception, in that the BACC for LAW improved by 29%. Based on these results, it can be deduced that RW not only increases fairness in outcomes, but, in the case of imbalanced data, enhances model performance measured by BACC. Moreover, the loss of accuracy is minor, as emphasized by the low percentages.

Table 7. LRC results after reweighing the training set. Performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. The percentages in brackets display the differences compared to the baseline results as presented by Table 4.

	ACC	BACC	F1	SPD	AOD
COMPAS	0.655 (−2.96%)	0.656 (−1.94%)	0.699 (−1.96%)	0.043 (+116.48%)	0.076 (+133.04%)
LAW	0.950 (−0.21%)	0.731 (+29.15%)	0.974 (−0.1%)	−0.005 (+91.8%)	0.014 (+111.2%)

However, the results reported in Table 7 are potentially biased, since the test sets were not modified accordingly. To ensure comparability Table 8 additionally displays the change compared to a baseline LRC after reweighing both training and test sets.

Table 8. LRC results after reweighing the training and test set. Performance was evaluated on ten repeated train-test splits. For SPD and AOD. values closer to zero indicate more fairness. The percentages in brackets display the differences compared to a baseline LRC evaluated on equal train-test splits.

	ACC	BACC	F1	SPD	AOD
COMPAS	0.658 (−2.08%)	0.651 (−2.11%)	0.700 (−1.82%)	0.094 (+135.47%)	0.093 (+139.57%)
LAW	0.0.951 (−0.11%)	0.535 (−4.46%)	0.975 (=)	0.001 (+101.72%)	0.015 (+113.16%)

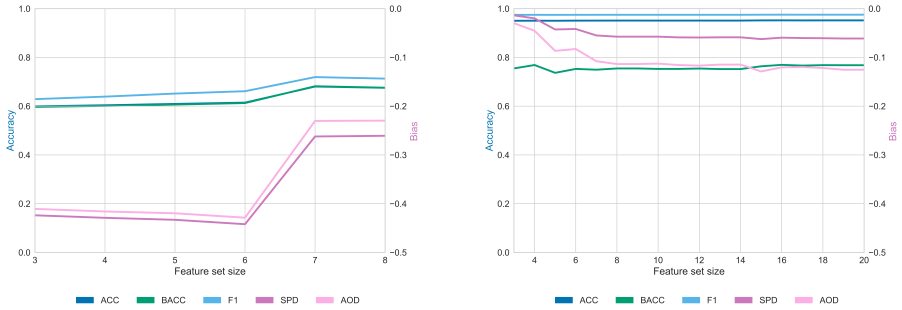
The results achieved by transforming the entire data sets did not differ much from before. Nevertheless, for LAW accuracy measured by BACC decreased, whereas adjusting only the training data positively impacted BACC. These differences were due to the fact that the stratified option for splitting was not available for the provided functions by the *AIF360* toolkit and, hence, the effect of the unbalanced class labels was even stronger compared to results by stratified 10-fold CV. To conclude, reweighing the binary target classes with respect to the protected feature contributes to fair outcomes. However, accuracy slightly decreases. The negative impact on accuracy is minor compared to the increase of fairness in outcomes.

7.5. Feature Set Size

Based on [11], feature set size can be defined as knowledge available about an individual, with each feature presenting his or her characteristics. The more features involved, the more information available about an individual, enhancing the probability of assessing the individual's ability correctly.

Figure 4 illustrates the behavior for accuracy and fairness for different counts of features. The order of the features in each set corresponds to the original order of the data sets, as given in the data pre-processing part under Section 6. For COMPAS, increasing the number of features enhanced both fairness and accuracy. In particular, the seventh feature, defined as *priors_count*, led to a significant increase in fairness and accuracy. In contrast, increasing the size of the feature set caused a loss of fairness for LAW. The loss in fairness remained almost constant from a feature set size of seven. Consequently, the first seven features, illustrating the performance scores of students, appeared to be biased. The bias was related to the protected feature whereby previous student performances evaluated by teachers are potentially subject to discrimination. The predefined accuracy measures slightly changed throughout different feature set sizes for LAW. Since some of the features

appeared to be more biased than others, based on the previous findings, the order of features additionally needed to be examined critically to identify whether the positive effect highlighted by [11] was due to the size of the sets or whether it was related to certain features.

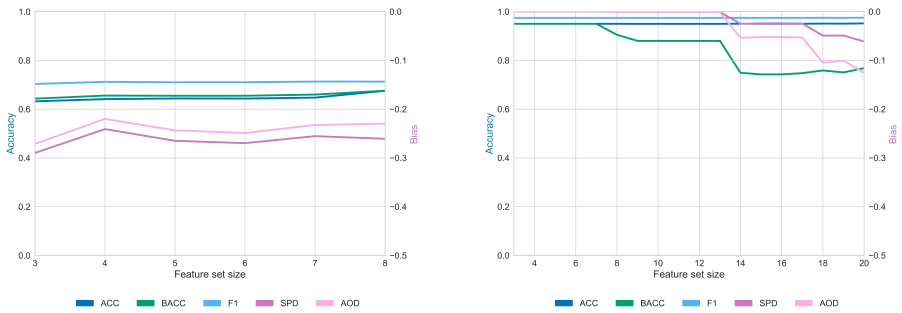


(**a**) (**b**)

Figure 4. Performance results for LRC on different feature set sizes. Model performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. (**a**) COMPAS; (**b**) LAW.

The new feature order for COMPAS is as follows: {1. race, 2. priors_count, 3. sex, 4. juv_other_count, 5. juv_misd_count, 6. c_charge_degree_M, 7. juv_fel_count, 8. age }; and for LAW: {1. race, 2–6. family_income, 7. part_time, 8–13. cluster_tier, 14. weighted_lsat_ugpa, 15. sex, 16. lsat, 17. ugpa, 18. zfygpa, 19. DOB_yr, 20. zgpa}.

The protected feature *race* was always set to be the first feature to make sure it was included throughout the iterative process. If *priors_count*, representing the number of prior crimes an individual committed in the past, was included in the first set of features studied for COMPAS, fairness was not further improved with increasing feature set size, as illustrated by Figure 5. The results for LAW also supported these findings: fairness decreased with the introduction of student performance variables like *weighted_lsat_ugpa*. Based on these findings, it can be concluded that the impact on fairness and accuracy is not related to the size of the feature sets, but rather to the information certain features add to the model. For example, in the COMPAS dataset, the feature "priors_count" significantly increases fairness, whereas in the LAW dataset, the feature "weighted_lsat_ugpa" decreases fairness. These features demonstrate high importance in their corresponding dataset, and, therefore, have a significant impact on model fairness. However, such important features are data-set specific and their exact impact needs to be studied in future work.



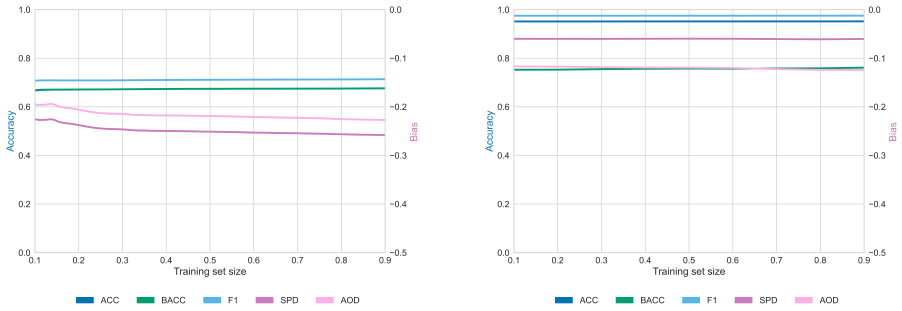
(**a**) (**b**)

Figure 5. Performance results for LRC on various feature set sizes of different orders. Model performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. (**a**) COMPAS; (**b**) LAW.

7.6. Training Set Size

In general, accuracy is expected to increase when more training data is used. However, if the training data is biased, the model becomes even better in depicting discriminating patterns. Therefore, the effect of training set size on fairness is expected to be negative, as claimed by [11]. Figure 6 represents the behavior of accuracy and fairness according to the size of the training set.

The results differed for both considered data sets. For COMPAS, adding more data points to the training set led to a small decrease in fairness, while accuracy was kept constant. On the other side, for LAW increasing training data appeared to have no significant effect on the analyzed instances. For both data sets considered, accuracy was not improved for enlarged training data.

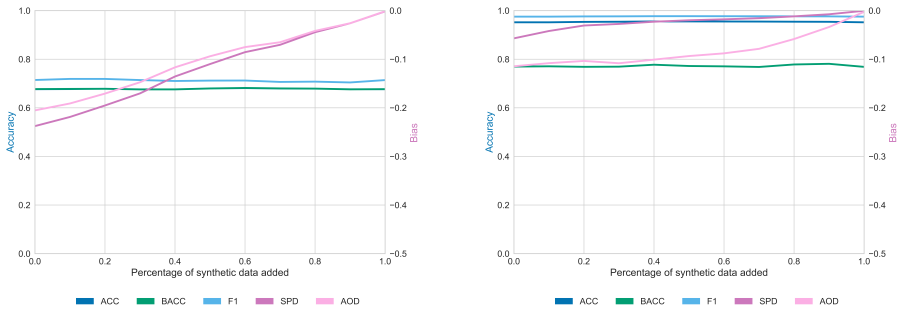


(**a**) (**b**)

Figure 6. Performance results for LRC on differently sized training sets. Model performance was evaluated using the average of ten stratified train–test splits. For SPD and AOD, values closer to zero indicate more fairness. (**a**) COMPAS; (**b**) LAW.

7.7. Data Augmentation

Ref. [10] claimed that adding realistic synthetic data points to the original set balances the protected feature classes and leads to a significant increase in fairness, while accuracy remains constant. The graphs displayed in Figure 7 illustrate the behavior of accuracy and fairness for percentage-wise added synthetic data. The start of the x-axis indicates no synthetic data points added to the original data set and the counterpart corresponds to 100% of data augmentation, meaning that the whole synthetic data set was added to the original data.



(**a**) (**b**)

Figure 7. Performance results for the LRC on augmented data. Synthetic data points are added to the training and test sets for evaluation. Performance was evaluated using stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. (**a**) COMPAS; (**b**) LAW.

By reversing the protected feature labels for the synthetic data sets and adding the most realistic data points to the original data sets, the model can be trained on larger training data where the racial groups are balanced. This additionally explains the result that fairness

maximization was reached for 100% data augmentation [10]. Accuracy was rather constant during the process. Table 9 displays the results compared to the baseline LRC.

Table 9. Performance results for LRC on augmented data, performing stratified 10-fold CV. For SPD and AOD, values closer to zero indicate more fairness. The results correspond to maximal fairness achieved by data augmentation based on [10]. The percentages in brackets display the differences compared to the baseline results as presented by Table 4.

	Data	ACC	BACC	F1	SPD	AOD
COMPAS	Train + Test	0.676 (+0.15%)	0.676 (+1.05%)	0.714 (+0.14%)	0.001 (+100.38%)	0.002 (+100.87%)
	Train	0.674 (−0.15%)	0.673 (+0.60%)	0.704 (−1.26%)	−0.225 (+13.79%)	−0.193 (+16.09%)
LAW	Train + Test	0.952 (=)	0.768 (+35.69%)	0.975 (=)	0.000 (+100%)	−0.003 (+97.6%)
	Train	0.952 (=)	0.762 (+34.63%)	0.975 (=)	−0.059 (+3.28%)	−0.116 (+7.2%)

Compared to the baseline performance results, data augmentation increased accuracy, especially for BACC from LAW when augmenting with 100% of the synthetic data points. Due to the flipping of the protected features and keeping the class labels unaltered, the groups were balanced. Consequently, the model had more data to be trained on to make more accurate predictions. However, since the test sets in ML were used for evaluation and, therefore, kept unaltered by default, synthetic data points were only added to the training sets as follows, while the test sets remained untouched.

For COMPAS, the training sets were enlarged by the entire synthetic data set, whereas for LAW 40% of the synthetic data points were added to maximize fairness. For both data sets, accuracy, measured by ACC, and the F1 score were not affected or slightly decreased by adding synthetic data points to the training sets and evaluating the model on unaltered test data. Yet the described method had a positive impact on BACC, especially with respect to LAW. This effect can be explained by balancing the privileged and unprivileged groups. The considered measures for fairness became less negative, but the difference compared to the baseline results was only small. In summary, performing data augmentation provides promising results with respect to fairness and accuracy. In both contexts analyzed, fairness is enhanced significantly without decreasing accuracy. If the test data is unaltered, the results do not differ significantly from the baseline results achieved by the LRC. However, the described approach accounts for imbalanced data increasing BACC significantly for LAW.

8. Discussion

Based on the majority of research, enhancing fairness always leads to a loss of accuracy [4]. More drastically, some researchers consider this trade-off to be unavoidable [3]. However, as [12] states, increasing accuracy might as well improve fairness due to minimizing the number of errors produced by the model. The results achieved by tuning the LRC support this statement. Optimizing the LRC with respect to ACC and BACC for balanced data led to an increase in fairness measured by SPD and AOD of about 1%. Modifying the classification threshold with respect to ACC for imbalanced data even led to an increase of the SPD score of 8%. In contrast, adjusting the considered classifier with respect to BACC appeared to have a major negative effect on fairness. Therefore, increasing accuracy for balanced data sets can be assumed to positively impact fairness in outcomes, whereas the impact on imbalanced data needs to be further investigated. Furthermore, adapting the classification threshold to maximize accuracy, measured by the F1 score, improved fairness significantly. For both data sets considered, the values achieved for SPD and AOD increased by, on average, 30%. However, accuracy measured by ACC and BACC decreased by 2–5%. To conclude, adapting the discrimination threshold appears to improve fairness.

The considered pre-processing mitigation in this study was RW. When training data was reweighed, based on protected features and labels, while the test set remained unchanged, there was a small decrease in accuracy, but a significant improvement in fairness. Specifically, SPD and AOD increased by 91–133%, while the loss of accuracy amounted to

0.1–3%. That is, reweighing the samples with respect to the protected feature and target class labels improves fairness significantly, whereas the loss of accuracy is only minor.

Furthermore, general ML practices, such as the size of the feature and training set, were analyzed. Changing the order of features added to the iterative process of evaluating different feature set sizes in displaying disparate behaviors of fairness and accuracy. Therefore, more features used do not intrinsically cause an increase in fairness and accuracy. Increasing the number of training samples did not appear to impact accuracy regarding COMPAS and LAW, either. In contrast, the approach of data augmentation, as proposed by [10], increases fairness and accuracy. Additionally, the findings of our work support the recommendation from [56] to evaluate model performance on several training and test sets to ensure performance stability. The results regarding fairness metrics are sensitive towards input variations.

As the various approaches emphasize, fairness and accuracy may even benefit each other. Furthermore, the results stress that the loss of accuracy is minor compared to the significant increase of fairness for the considered pre-processing mitigation strategy and the approach of data augmentation. To account for biases, new biases could be created by adding more data or reweighing samples to ensure fairness in outcomes. The results also indicate that the approaches used to improve fairness do not necessarily lead to a loss of accuracy, but the difference in distributions between the training and test sets does. Limitations evolved during our analysis due to the choice of data sets, setting of model training and performance evaluation. The majority of results illustrate the average performance results for stratified 10-fold CV. Changing this setting might lead to disparate findings. All approaches explored were tested on one balanced and one imbalanced data set. They further apply to the stage of data collection and focus on either data balancing by RW, feature selection or enlarging the size of the data set. In addition, the only bias mitigation pre-processing strategy included in the analysis was the RW approach invented by [21].

Further research may be aimed at filling all these gaps, by involving other data sets, exploring other stages of the ML pipeline, in addition to examining other fairness definitions [17] and related measures [16]. Other measures can also be used to define the "cost of fairness" [4]. This is also true for classifiers and performance stability. Even though various approaches were employed to analyze the relationship between accuracy and fairness, the approaches were limited because they only referred to the first stage of the ML pipeline. All approaches explored focused on data balancing by RW, feature selection or enlarging the size of the data set, and applied to the stage of data collection. In addition, the bias mitigation pre-processing strategy included in the analysis was the RW approach invented by [21]. To maximize accuracy and simultaneously study the effect on fairness, optimization was performed before modeling by setting hyperparameters with respect to the predefined accuracy metrics. However, the relationship of accuracy and fairness is not limited to this stage of the ML pipeline.

Moreover, as emphasized throughout this work, the definitions of fairness and accuracy play a major role when assessing their relationship. Depending on the selection of performance measures the results can vary significantly. Fairness in the context of ML has no universal definition [17]. The selection needs to be made with respect to: (1) the context of the analysis and (2) the given characteristics of the data set. [16] recommends reporting on at least two fairness measures. Furthermore, the approaches were tested on one balanced and one imbalanced data set. In particular, the findings for the imbalanced data set should be verified on further examples to enhance external validity.

9. Conclusions

Several strategies can be used to improve the fairness of ML models. First, increasing the accuracy of models for balanced data sets can have a positive impact on fairness. Second, adjusting the classification threshold to maximize accuracy, as measured by the F1 score, can also improve fairness to a large extent. In addition, it was shown that bias mitigation

techniques, such as reweighing samples with respect to protected characteristics and target class designations, can significantly improve fairness with minimal loss of accuracy. Finally, data augmentation is another approach that can be used to improve fairness.

Our findings also underscore the fact that there is no universal definition for either fairness or accuracy, and both need further investigation in different application domains. It is not only the choice of fairness metrics that matters, but also the definition of model correctness. Instead of focusing on the "cost of fairness" [4], further studies should stress the role of simultaneously choosing a definition of accuracy. Furthermore, studies on human preferences of metrics across cultural contexts and understanding of all metrics are necessary.

Author Contributions: Conceptualization, A.L., S.-C.M., T.E. and B.F.; methodology, A.L., T.E. and B.F.; software, A.L.; validation, S.-C.M. and T.E.; writing—original draft preparation, A.L.; writing—review and editing, S.-C.M., T.E. and B.F.; supervision, T.E. and B.F.; project administration, B.F. All authors have read and agreed to the published version of the manuscript.

Funding: Co-funded by the European Union. The pro_digital European Digital Innovation Hub (EDIH) at Technical University of Applied Sciences Wildau received co-funding from the European Union's DIGITAL EUROPE Programme research and innovation programme grant agreement No. 101083754.

Data Availability Statement: The data sets used in this research are publicly available, see Section 6.1.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. COMPAS data set features adapted from [60].

Feature Name	Type	Value	Description
race	Binary	{African-American, Caucasian}	Race
sex	Binary	{Male, Female}	Gender
age	Numerical	[18–96]	Age in years
juv_felt_count	Numerical	[0–20]	Juvenile felony count
juv_mids_count	Numerical	[0–13]	Juvenile misdemeanor count
juv_other_count	Numerical	[0–17]	Juvenile other offenses count
priors_count	Numerical	[0–38]	Prior offenses count
c_charge_degree	Binary	{F, M}	Charge degree of original crime
two_year_recid	Binary	{0, 1}	Rearrested within two years

Table A2. LAW data set features adapted from [60].

Feature Name	Type	Value	Description
race	Binary	{Non-White, White}	Race
gender	Binary	{Female, Male}	Gender
lsat	Numerical	[11–48]	The student's LSAT score
ugpa	Numerical	[1.5–4]	The student's undergraduate GPA
zfygpa	Numerical	[−3.35–3.48]	The first year law school GPA
DOB_yr	Numerical	[10–71]	Year of birth
zgpa_felt_count	Numerical	[−6.44–4.01]	The cumulative law school GPA
family_income	Categorical	5	The student's family income bracket
part_time	Binary	{0,1}	Working full- or part-time
cluster_tier	Categorical	6	Tier of school
weighted_lsat_ugpa	Numerical	[288.95–1000]	Weighted LSAT and UGPA score
pass_bar	Binary	{0, 1}	Passed bar exam on first try

References

1. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Broadway Books: New York, NY, USA, 2016.
2. Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.* **2017**, *31*, 1060–1089. [CrossRef]
3. Cooper, A.F.; Abrams, E.; Na, N. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Online, 19–21 May 2021; pp. 46–54.
4. Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq, A. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 797–806.
5. Menon, A.K.; Williamson, R.C. The cost of fairness in binary classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 107–118.
6. Zhao, H.; Gordon, G. Inherent tradeoffs in learning fair representations. *J. Mach. Learn. Res.* **2022**, *23*, 2527–2552.
7. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S. On the (im) possibility of fairness. *arXiv* **2016**, arXiv:1609.07236.
8. Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.Y.; Liu, S.; Varshney, K. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 2803–2813.
9. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* **2021**, *50*, 3–44. [CrossRef]
10. Sharma, S.; Zhang, Y.; Ríos Aliaga, J.M.; Bouneffouf, D.; Muthusamy, V.; Varshney, K.R. Data augmentation for discrimination prevention and bias disambiguation. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–8 February 2020; pp. 358–364.
11. Zhang, J.M.; Harman, M. "Ignorance and Prejudice" in Software Fairness. In Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), Madrid, Spain, 25–28 May 2021; pp. 1436–1447.
12. Hellman, D. Measuring algorithmic fairness. *Va. Law Rev.* **2020**, *106*, 811–866.
13. Wick, M.; Tristan, J.B. Unlocking fairness: A trade-off revisited. In Proceedings of the Advances in neural information processing systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
14. Fish, B.; Kun, J.; Lelkes, Á.D. A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, FL, USA, 5–7 May 2016; pp. 144–152.
15. Rodolfa, K.T.; Lamba, H.; Ghani, R. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nat. Mach. Intell.* **2021**, *3*, 896–904. [CrossRef]
16. Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In Proceedings of the Conference Fairness, Accountability and Transparency, New York, USA, 23–24 February 2018; Volume 1170, p. 3.
17. Verma, S.; Rubin, J. Fairness definitions explained. In Proceedings of the 2018 IEEE/ACM International Workshop on Software Fairness (Fairware), Gothenburg, Sweden, 28–29 May 2018; pp. 1–7.
18. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *Acm Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
19. Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 339–348.
20. Barocas, S.; Hardt, M.; Narayanan, A. Fairness in machine learning. *NIPS Tutor.* **2017**, *1*, 2.
21. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
22. Dunkelau, J.; Leuschel, M. *Fairness-Aware Machine Learning—An Extensive Overview*; Universität Düsseldor: Düsseldorf, Germany, 2019.
23. Larson, J.; Mattu, S.; Kirchner, L.; Angwin, J. How we analyzed the COMPAS recidivism algorithm. *ProPublica* **2016**, *9*, 5.
24. Besse, P.; del Barrio, E.; Gordaliza, P.; Loubes, J.M.; Risser, L. A survey of bias in machine learning through the prism of statistical parity. *Am. Stat.* **2021**, *76*, 188–198. [CrossRef]
25. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
26. Kim, J.S.; Chen, J.; Talwalkar, A. FACT: A diagnostic for group fairness trade-offs. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 5264–5274.
27. Zhang, J.M.; Harman, M.; Ma, L.; Liu, Y. Machine learning testing: Survey, landscapes and horizons. *IEEE Trans. Softw. Eng.* **2020**, *48*, 1–36. [CrossRef]
28. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012; pp. 214–226.
29. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.

30. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef]
31. Grgic-Hlaca, N.; Zafar, M.B.; Gummadi, K.P.; Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In Proceedings of the NIPS Symposium on Machine Learning and the Law, Barcelona, Spain, 5–10 December 2016; Volume 1, p. 2.
32. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
33. Suresh, H.; Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization, New York, NY, USA, 5–9 October 2021; pp. 1–9.
34. Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C. Learning fair representations. In Proceedings of the International Conference on Machine Learning, Miami, FL, USA, 4–7 December 2013; pp. 325–333.
35. Bellamy, R.K.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* **2018**, arXiv:1810.01943.
36. Mellin, W. Work with new electronic 'brains' opens field for army math experts. *Hammond Times* **1957**, *10*, 66.
37. Babbage, C. Chapter VIII—Of the analytical engine. In *Passages from The Life of a Philosopher*; Longman: London, UK, 1864; pp. 112–141.
38. Calders, T.; Žliobaitė, I. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 43–57.
39. Suresh, H.; Guttag, J.V. A framework for understanding unintended consequences of machine learning. *arXiv* **2019**, arXiv:1901.10002.
40. Danks, D.; London, A.J. Algorithmic Bias in Autonomous Systems. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia, 19–25 August 2017; Volume 17, pp. 4691–4697.
41. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 59–68.
42. Silva, S.; Kenney, M. Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon* **2018**, *55*, 9–37.
43. Kamiran, F.; Calders, T. Classifying without discriminating. In Proceedings of the 2009 2nd International Conference on Computer, Control and Communication, Karachi, Pakistan, 17–18 February 2009; pp. 1–6.
44. Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; Varshney, K.R. Optimized pre-processing for discrimination prevention. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
45. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 10–13 August 2015; pp. 259–268.
46. Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; Zemel, R. The variational fair autoencoder. *arXiv* **2015**, arXiv:1511.00830.
47. Hajian, S.; Domingo-Ferrer, J.; Martinez-Balleste, A. Rule protection for indirect discrimination prevention in data mining. In Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence, Changsha, China, 28–30 July 2011; pp. 211–222.
48. Edwards, H.; Storkey, A. Censoring representations with an adversary. *arXiv* **2015**, arXiv:1511.05897.
49. Kozodoi, N.; Jacob, J.; Lessmann, S. Fairness in credit scoring: Assessment, implementation and profit implications. *Eur. J. Oper. Res.* **2022**, *297*, 1083–1094. [CrossRef]
50. Noriega-Campero, A.; Bakker, M.A.; Garcia-Bulle, B.; Pentland, A. Active Fairness in Algorithmic Decision Making. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), Honolulu, HI, USA, 27–28 January 2019; pp. 77–83. [CrossRef]
51. Wightman, L.F. *LSAC National Longitudinal Bar Passage Study*; Law School Admission Council: Newtown, PA, USA, 1998.
52. Kleinbaum, D.G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M. *Logistic Regression*; Springer: New York, NY, USA, 2002.
53. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
54. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Gradient Boosting for Classification. 2011. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html (accessed on 12 May 2022).
55. Bengfort, B.; Bilbro, R. Yellowbrick: Visualizing the scikit-learn model selection process. *J. Open Source Softw.* **2019**, *4*, 1075. [CrossRef]
56. Friedler, S.A.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.P.; Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 329–338.
57. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Cross-Validation: Evaluating Estimator Performance. 2011. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed on 12 May 2022).
58. Hoffman, S.C. The AIF360 Team Adds Compatibility with Scikit-Learn. 2020. Available online: https://developer.ibm.com/blogs/the-aif360-team-adds-compatibility-with-scikit-learn/ (accessed on 12 May 2022).

59. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*; MIT Press: Cambridge, MA, USA, 2018.
60. Quy, T.L.; Roy, A.; Iosifidis, V.; Ntoutsi, E. A survey on datasets for fairness-aware machine learning. *arXiv* **2021**, arXiv:2110.00530.